AUTOMATING MOLECULAR FORCEFIELD DEVELOPMENT AND IMPROVING PARAMETERIZATION METHODS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Gregory Michael Casee, Jr.

August 2019

Abstract

At the heart of all Molecular Dynamics simulations lies an energy potential that seeks to capture the underlying quantum mechanical interactions between atoms and molecules. However, describing these interactions is difficult, and developing parameters and functional forms for molecular forcefields is a major roadblock for researchers looking to study new systems computationally. We have simplified forcefield parameterization through an iterative optimization approach that automatically generates new training sets, calculates their energies using Density Functional Theory, and fits the results to the desired potential. We also implement a method of joining distinctive forcefields that allows for flexible choices of short- and long-range potentials that efficiently model reactive environments. This method is illustrated by modeling lead-sulfide quantum dots and their passivating ligands. Finally, we propose methods of weighting training set data based on geometry and energy considerations to increase model accuracy during simulations.

Biographical Sketch

Greg Casee graduated from Lehigh University in 2017 with a Bachelor of Science in Chemical Engineering. During his time at Lehigh, Greg worked on several research projects in areas such as developmental biology, adhesives and nanoparticle coatings, and additive manufacturing of concrete. He discovered computational chemistry during a simulations class in his senior year and worked on a project in that area during the following summer. Greg joined the Chemical and Biomolecular Engineering graduate program at Cornell in the fall of 2017 to pursue research in this field and intends on further developing his computational research skills beyond graduate school.

To Ryan, Emily, Brooke, and Brett, may you never have to write a thesis as it is a long and terribly boring endeavor.

To Professor Rangarajan and Professor Mittal, for introducing me to the world of computational chemistry.

And finally, to StackOverflow, for teaching me how to code.

Acknowledgments

I would like to acknowledge a few individuals who made this work possible:

First, Dr. Henry Herbol was instrumental to my research success. Henry was a huge help to me as I was getting started within the group. He has also been a great mentor for research and computer science. I am also grateful to the rest of the Clancy lab members as they have all contributed to this work through their feedback and support.

I would also like to thank Professor Hanrath and Professor Stroock for stepping in to be my committee members so that I may defend the research I have done at Cornell.

Finally, I am grateful for all of the advice and support from numerous professors in helping me navigate some unforeseen issues that arose along the way.

Contents

List of Tables

List of Figures

DFT Calculations of Quantum Dot Systems

1.1 Introduction to Quantum Dots

Although quantum dots (also described as nanocrystals or nanoparticles) were first synthesized in 1981, these semiconducting nanoparticles have only recently garnered research interest to use them as tunable building blocks for larger nanostructures [1]. Quantum dots are particles with diameters typically on the order of tens of nanometers that exhibit discrete electronic states, similar to those of atoms or molecules [2]. The size-dependent quantum confinement of quantum dots allows for tunable electronic and optical properties by varying the diameter or composition during synthesis. In particular, quantum dots made from lead and a chalcogenide such as sulfur or selenium are currently widely studied. Advancements in synthesis allow these quantum dots to be made in solution in a single pot with high fidelity—on the order of 4 to 10 nm in diameter with dispersions as small as 4 percent [3]. The ability to be manufactured in solution with high control over the process provides an economical path to scale up the technology and, as a result, has attracted much commercial interest.

Since the composition, size, and shape of quantum dots can be precisely controlled during synthesis, quantum dots are ideal building blocks for larger structures. Through oriented attachment and fusion along crystal facets, the nanocrystals can form 1D nanowires, 2D superlattices, or larger structures with varying degrees of long range order [4]. Synthesizing nanostructures from tunable building blocks will potentially unlock a new class of materials with designer properties. In particular, 2D superlattices may have applications in photovoltaics or optoelectronics due to their strong electronic coupling between dots and

predicted high efficiency of electron and hole transport [4–6]. These properties make quantum dots a promising candidate for next generation photovoltaics that could boast higher efficiencies than current silicon-based technologies [7]. Additionally, quantum dot superlattice construction leaves well-defined nanopores whose size can be controlled by the size of the constituent dots. The natural surface charge of the crystalline facets allows for selective permeability to either cations or anions in a manner similar to other single-layer materials [8]. Furthermore, doping these membranes with catalytically active atoms while carefully designing the selectivity of the nanopores could create structures that combine reaction with separation in a single material. The ability to tune the properties of 2D structures by changing the characteristics of the nanocrystal building blocks will lead to new class of designer materials for photovoltaics, separations, and catalysis. Although individual quantum dots can be synthesized to high precision, fusing dots to form higher order structures still suffers from local defects that are detrimental to creating long-range order. Further exploration of the fundamental science of superlattice construction needs to be done before these materials are experimentally accessible.

The process of converting individual colloidal quantum dots to colloidal quantum dot assemblies and finally to epitaxially connected quantum dot solids involves many interconnected processes. Superstructures are often formed at a liquid-liquid interface between immiscible solvents to promote two-dimensional film growth. After synthesis, colloidal quantum dots—particles with passivating ligands that both protect the crystal surface and promote solubility—are dissolved in a nonpolar solvent, such as hexane, and are deposited on top of an immiscible ethylene glycol subphase. As the solvent is evaporated, colloidal quantum dots begin to self-assemble at the interface to form hexagonal colloidal quantum dot assemblies [9]. At this point, the quantum dots are still passivated by ligands and are not yet epitaxially attached; an interplay of enthalpic and entropic effects drive the suspended particles to undergo a phase transition from a hexagonal to a square lattice formation as more solvent is driven off [10]. To induce epitaxial bond formation, a chemical trigger—typically compounds

with amines—is added to the system which strips the ligands off of the quantum dot, exposing "sticky" facets which then allow the dots to bond. [11]. Unfortunately, the resultant superlattice suffers from several key defects that negatively impact the performance of the material. Currently, grain sizes of polycrystalline quantum dot solids are limited to about 5 microns; however, on the atomic scale, reconstructed TEM images show that epitaxial connections vary in size, and the dots themselves vary in shape [12]. These defects destroy any short range order on the atomic scale, even though the superlattice demonstrates long range order across microns.

The surfaces of colloidal quantum dots are passivated by ligands that stabilize the nanoparticle, make it soluble in hydrocarbon solvents, and prevent adjacent quantum dots from fusing prematurely. These ligands are typically oleic acid left over from the quantum dot synthesis, but can be exchanged to alter the electronic properties of the colloidal quantum dots [13, 14]. The length of these ligands, as well as the ligand density on the nanocrystal surface, affect the structure of the final superlattice by changing the way colloidal quantum dots arrange in solution [15, 16]. However, the ligands must be stripped from the quantum dot, exposing the bare facets, before irreversible epitaxial inter-dot bonds can form. The chemical trigger binds to passivating ligands, diffuses away from the colloidal quantum dot assembly, and dissolves into the subphase. However, the exact mechanism by which the ligands are removed is not well known. Ligands are constantly in exchange with the surrounding environment, and the chemical trigger may react with these ligands in solution, barring them from reattaching to the quantum dot surface [11]. Alternatively, the chemical trigger could actively strip passivating ligands from the surface by adsorbing onto the quantum dot, binding to a ligand, and finally desorbing as a larger complex. Understanding the precise role that the chemical trigger plays in inciting quantum dot fusion will aid in the design of different triggers or experimental conditions to improve material quality.

1.2 DFT Methods for Quantum Dot Systems

Even the smallest experimentally relevant quantum dots of about 3 nm contain over 400 atoms, not including any ligands or solvent. The size of these systems excludes the use of density functional theory (DFT) or other ab initio methods to study entire quantum dots. However, we made approximations and divided the system into manageable pieces that are small enough to study with accurate DFT methods.

The first simplification I made is to shorten the lead oleate chains. One oleate chain contains 18 carbon atoms, which not only add a significant number of electrons to the self consistent field calculations, but also contribute many degrees of freedom for geometry optimizations. Fortunately, binding energies of lead carboxylates are insensitive to the number of carbon atoms in the chain, justifying the use of a shorter hydrocarbon. For example, I calculated the binding energy of lead acetate and lead pentanoate to ethylenediamine and found that they differ by only 0.005 kcal/mol (-18.326 and -18.321 kcal/mol, respectively). Therefore, for all subsequent DFT calculations, the ligands are assumed to be lead acetate (2 carbons) as an approximation to longer lead oleate molecules (18 carbons), and the chemical trigger studied is ethylenediamine (EDA).

The ligand chemistry of quantum dots was studied using the PBE functional which is widely employed in the literature to study these systems [17]. Two software packages were used to perform the DFT calculations: Quantum Espresso and Orca. Quantum Espresso is an open-source periodic DFT solver that is useful for probing the surface of nanocrystals or other naturally periodic systems [18]. The periodic DFT calculations utilized projector augmented wave method (PAW) pseudopotential files [19][20]. Orca is a free DFT package that is suited for exploring ligand reactions in solution [21]. All reported calculations were done with the default 2 valence triple-zeta polarization (def2-TZVP) basis set with an effective core potential on the lead atoms as needed [22]. To mimic experimental conditions, the reaction energies were calculated using the COSMO implicit solvation model for hexane [23].

1.3 Ligand Interactions with Quantum Dot Surface

DFT calculations provide an overview of the energy landscape for ligand reactions and can find the relative binding energies for ligands attached to quantum dot surfaces. Since quantum dots fuse along the 100 facets, I explored the binding energies of the relevant species to this facet. I first had to determine where the ligand affixes itself on the 100 surface since there are 3 favorable positions: placed on top directly above the chalcogen, forming a bridge between lead and selenium, and in a hollow at the center of four atoms. It appears that ligands favor the hollow position (at the center of four surface atoms on the 100 surface). These energies were estimated with a gamma point calculation, and for a neutral lead atom ligand, the bridge and hollow sites are 0.67 and 1.34 eV more stable than in the the top configuration, respectively.

The binding energies for lead acetate ($Pb(OAc)_2$), ethylenediamine (EDA), and a geometry optimized complex of the two were calculated on a PbSe 100 surface at a ligand density of 0.65 lig/nm^2. Here, the binding energy is defined to be the energy difference between the free and bound species, that is, the energy to remove a molecule from the surface. The results are summarized in Table 1.1. The magnitudes of these energies match those from larger scale simulations that predict a binding energy for lead oleate at the same ligand density of about 12 kcal/mol [24]. My periodic DFT calculations show that the chemical trigger, ligand, and complex all have similar bonding strengths to PbSe 100 surfaces. The addition of the chemical trigger does not appear to weaken the bond between the quantum dot surface and the passivating ligand; in fact, the entire complex appears to bind more strongly to the surface than the ligand does alone. Finally, the binding energy for EDA is only marginally less than that of lead acetate, suggesting that the facets may still be covered by the chemical trigger even after stripping the ligands off the surface. This result suggests that the quantum dot facets may not be as bare as originally thought, and that the chemical trigger may act as more than just a ligand stripping agent and play a role in quantum dot fusion.

Table 1.1: Binding Energy to PbSe 100 Surface

Species	Binding Energy [kcal/mol]
$Pb(OAc)_2$	10.3
EDA	8.9
$Pb(OAc)_2$+EDA Complex	12.1

1.4 Ligand Reactions in Solution

In addition to surface reactions, the ligands and the chemical trigger can bind in solution to form higher order complexes. For example, it has been shown that the dimerization of lead oleate ligands is an important process in the nucleation of lead sulfide quantum dots [25]. The aggregation of ligands with the chemical trigger might prevent the ligand from reabsorbing onto the crystal surface, which would describe how the surfaces become depassivated with the addition of ethylene diamine. A summary of the reactions of lead acetate (L) and ethylenediamine (E) and their energies are listed in Table 1.2.

Table 1.2: Reaction Energies for Aggregation of Lead Acetate (L) and Ethylenediamine (E)

Reaction	Energy [kcal/mol]
1L + 1E $\rightarrow$ 1L1E	-14.9
1E +1L1E $\rightarrow$ 1L2E	-6.7
1E +1L2E $\rightarrow$ 1L3E	-6.6
1L + 1L $\rightarrow$ 2L	-11.3
2L + 1E $\rightarrow$ 2L1E	-11.6
1E + 2L1E $\rightarrow$ 2L2E	-12.4
1E + 2L2E $\rightarrow$ 2L3E	-10.1
1E + 2L3E $\rightarrow$ 2L4E	-6.4

These results demonstrate that it is energetically favorable for multiple ligands and EDA molecules to bind together. However, there is a diminishing effect as more species are added

to the complex. Finally, these results predict that it is energetically favorable for the species to bind together, but larger scale methods are needed to determine rates and entropic effects.

Classical Forcefields and SMRFF

2.1 Introduction to Molecular Dynamics

Molecular dynamics (MD) is one method used to study large chemical systems up to the order of tens of thousands of atoms. Unlike density functional theory and similar methods that rely on quantum mechanics, molecular dynamics aims to describe atoms and molecules classically using Newton's equations of motion. Given a potential that describes how the atoms interact, MD solvers can set the particles into motion and track their trajectories over time, subject to imposed constraints, or ensembles. Some common ensembles include the canonical (constant number of particles, volume, and temperature) and the isothermal-isobaric (constant number of particles, pressure, and temperature) ensembles. Other thermodynamic contraints can be imposed on a molecular dynamics simulation depending on the properties that are being calculated. MD can be used to calculate bulk properties that are inaccessible by DFT such as diffusivities and densities, as well as provide insight into reaction and free energy pathways.

2.2 Introduction to Molecular Forcefields

Programs that solve Newton's equations of motion are straightforward to write, and plenty of open-source MD solvers are available [26][27]. However, the crux of performing accurate MD simulations is having accurate descriptions of how atoms interact, or a forcefield. Solving ballistic physics problems classically is relatively simple since bodies only interact through gravitational or coulombic interactions. Atoms differ in that they interact according to the laws of quantum mechanics, which requires solutions to the time-dependent Schrödinger equation. Unfortunately, solving this equation impossible analytically, and computationally expensive numerically. Therefore, classical forcefields need to be created that approximate exact energies from quantum mechanics without performing costly ab initio calculations.

Forcefields range from simple, two-body interactions to complex functions that hardly look

simpler than just solving the Schrodinger equation outright. As with classical forcefields, interatomic potentials can be written as functions of positions of the bodies, as described by Equation 2.1:

$$E = \sum_{i}^{N} V_1(r_i) + \sum_{i,j}^{N} V_2(r_i, r_j) + \sum_{i,j,k}^{N} V_3(r_i, r_j, r_k) + \dots \qquad (2.1)$$

in which V_1, V_2, V_3 are the one-, two-, and three-body terms, respectively, and the summations are over all N atoms in the system at positions r. The first term is only used in the case of an external field, such as an electric potential imposed on the entire system. The higher order terms are what need to be described by a potential. However, forcefields rarely go beyond three body terms because the number of unique combinations, and thus the number of parameters needed to describe the system, increases drastically.

Forcefields can be classified in a number of ways, but two of the major classifications are how many body interactions the field considers, and whether it is reactive or nonreactive. If atoms only interact pairwise, then the force a particular atoms experiences is dependent on the sum of all the individual interactions it has with all surrounding atoms. In contrast, three body interactions mean a third atom influences the potential felt between two other atoms. This can be imagined physically as a third body polarizing nearby atoms, and thus changing how they interact. Three- and four- body interactions are often realized in forcefields through angle and dihedral interactions, respectively.

Forcefields can also be divided into reactive and nonreactive. A nonreactive forcefield is one in which the bonds between atoms are specified ahead of time, and they cannot change throughout the simulation. In contrast, a reactive forcefield does not have predefined bonds, and atoms are allowed to move freely and change with what they are coordinated with.

One of the simplest yet commonly used two-body potentials is the Lennard-Jones potential (LJ) [28]. The functional form of the LJ potential is shown in Equation 2.2, and features a

long-range attractive term combined with a short-range repulsive term.

$$V_{LJ} = 4\epsilon\Big[\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big] \tag{2.2}$$

The LJ potential contains only two parameters, sigma and epsilon, which are a measure of distance and energy, respectively. The attractive r^6 term is meant to capture dispersion forces, while the repulsive r^{12} dominates at short distances and approximates the high energy of overlapping electron orbitals. The LJ potential can also be considered a reactive forcefield since no bonds are specified and atoms are free to disassociate. The LJ potential is often combined with an electrostatic potential for charged systems,

$$V_E = k\frac{q_i q_j}{r} \tag{2.3}$$

in which q_i and q_j are the charges of the two particles separated by a distance r. However, the LJ and electrostatic potentials are often only used to describe inter-molecular interactions in conjunction with other potentials that describe intra-molecular bonds.

Unlike the LJ potential, many forcefields explicitly define bonded interactions within molecules to capture the geometry of the species. One example is the Optimized Potentials for Liquid Simulations (OPLS), which was first developed to simulate protein structures and is now widely employed to study hydrocarbon systems [29][30]. Atoms within a molecule include up to four-body terms (dihedrals), while atoms in neighboring molecules interact only pair-wise (two-body interactions), but is considered non-reactive since atoms cannot change with what they are bonded. OPLS largely relies on a ball and spring model in which atoms are connected via a spring potential to mimic atomic bonds. In addition, spring-like potentials are added to enforce angles and dihedrals within the molecule. Finally, atoms in other molecules, or atoms separated in the same, large molecule, interact via LJ and coulombic

forces as described above.

A reactive forcefield similar to the LJ potential is the Morse potential [31]. The original formulation is reproduced below. Two of the parameters have a direct physical significance: D is the dissociation energy and r_0 is the minimum energy distance between two atoms. This function was proposed as a solution to the quantum harmonic oscillator that captures both the anharmicity of diatomic molecule vibrations as well as the ability to dissociate entirely.

$$V_{Morse} = De^{-2\alpha(r-r_0)} - 2De^{-\alpha(r-r_0)} \tag{2.4}$$

Finally, the Tersoff potential, named after the IBM research who proposed it, is a three-body, reactive potential [32]. The physical basis for such a potential is that all interatomic interactions are affected by their local environment. Tersoff employs pair-wise attractive and repulsive terms similar to those found in Morse, except that the strength of the attraction is modified by a term that captures the effect of a third body on the pair. This modification is a complex function of the third body's distance from the pair, as well as the angle it makes with the other two atoms. Although the Tersoff potential is more accurate than those that only calculate two-body interactions, it comes at a large computational cost of having to sum over all permutations of three atoms within the cutoff radius. Additionally, parameterizing a Tersoff potential is difficult since it features approximately 11 parameters per atom type in the system.

2.3 PbS Dot Example

As stated earlier, methods beyond DFT must be used in order to study entire quantum dots at experimentally relevant sizes. There are few computational studies on larger quantum dot systems using Molecular Dynamics (MD), but literature values for LJ parameters and atom charges can not reproduce DFT energies since they were fit solely to experimental lattice parameters and elastic constants of bulk crystals [33]. To rectify this, I trained a LJ and

Coulombic potential to describe pair-wise interactions between lead and sulfur atoms.

I used 32 atom mini-quantum dots with perturbed geometries as training sets to fit LJ parameters and charges for lead and sulfur atoms using global optimization methods that minimize the error between the classical forcefield and energies from DFT. The results of the fitting along with an example of the resulting dot dynamics are shown in Figure 2.1.

The error function was the sum of the square differences between DFT and classical energies divided by the DFT energy, as shown in Equation 2.5. Calculating the error in this manner biases the parameters to better match the DFT data at lower energies where the system will spend more time, and only roughly fit DFT data at higher energies. Without this weighting, the high energy systems that were randomly generated and used to fit the parameters skew the forcefield and create unstable dynamics. The optimized LJ parameters and Coulombic charges are reported below in Table 2.1.

$$Error = \sum \frac{(E_{DFT} - E_{FF})^2}{E_{DFT}} \tag{2.5}$$

Table 2.1: Optimized LJ parameters and atomic charges for PbS.

Atom	Charge	ϵ [kcal/mol]	σ [Å]
Pb	+1.31	0.0462	4.32
S	-1.31	0.0251	3.41

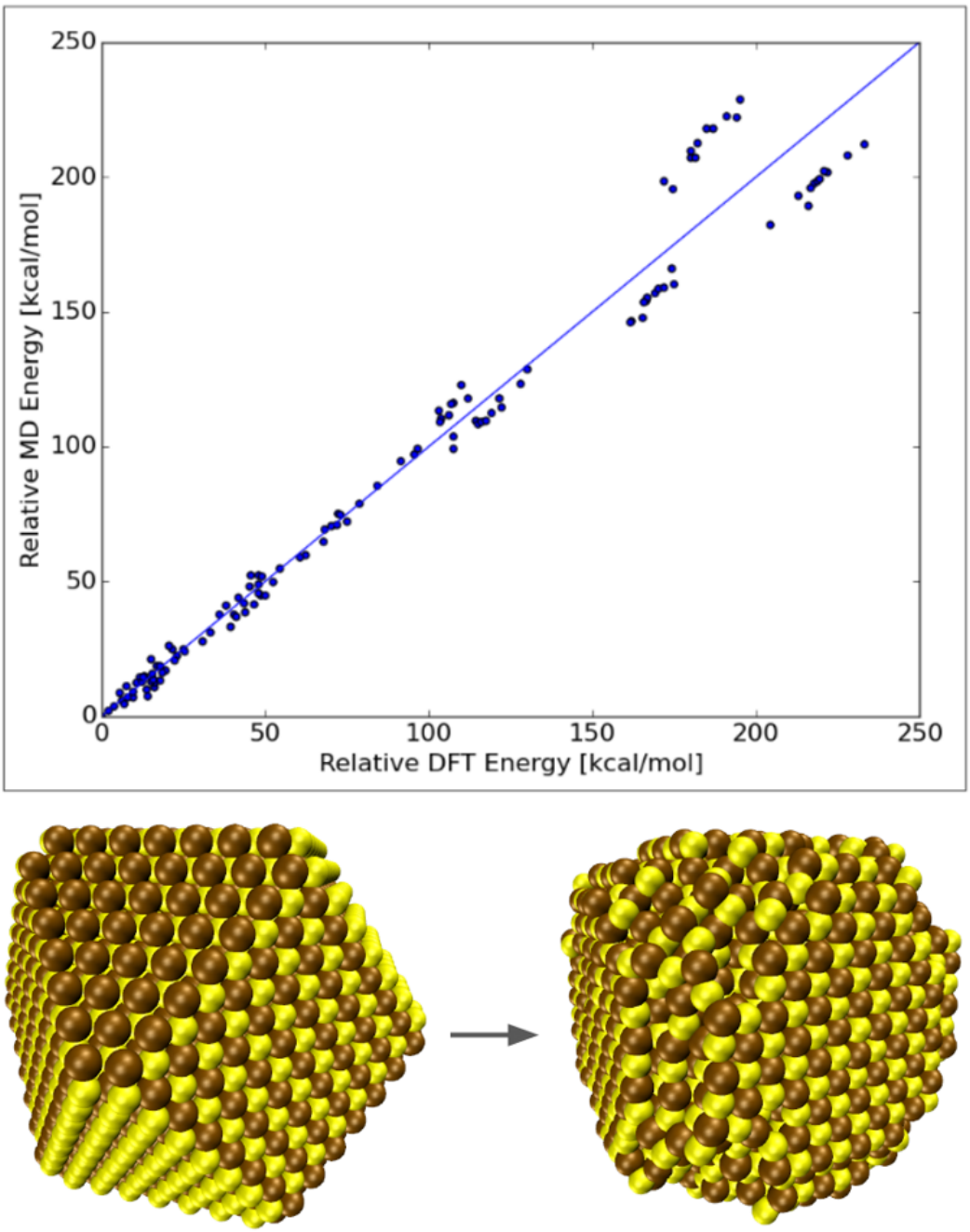


Figure 2.1: Above: Errors for PbS potential fitting; points on the drawn line indicate a perfect match between DFT and MD energies. Below: A perfectly octahedron quantum dot without passivating ligands will reconstruct to remove high energy 111 surfaces.

Molecular dynamics simulations of an approximately 4 nm PbS nanocrystal using these updated parameters exhibit restructuring of Pb or S terminated 111 facets to form more stable 100 surfaces. This behavior is consistent with both previous simulation data and experimental results [24, 34]. This successful test of the viability of using MD techniques and optimized forcefield parameters suggest that this approach can be used for further study of larger, experimentally relevant quantum dots. However, more complex descriptions are required to capture all of the interactions between the ligands, quantum dots, chemical trigger molecules, and solvent than can be described by a simple LJ potential.

2.4 Simple Molecular Reactive Force Field

In the Clancy lab, a new atomic potential dubbed SMRFF (Simple Molecular Reactive Force Field) has been developed and used to study the nucleation of lead sulfide systems [35]. SMRFF is a hybrid potential that can join together two different forcefields with a smooth transition between them. At close interatomic distances, atoms can interact via complex potentials such as Tersoff (for example), and then smoothly transition to simpler long-range potentials such as LJ and Coulomb. This approach results in an increase in simulation speed without a large loss in accuracy and thus allows for larger MD simulations to be conducted than those using a complicated forcefield alone.

Since force is simply the gradient of the potential, one requirement of all forcefields is that they be smooth, and thus differentiable over the domain. Since SMRFF joins two different potentials together, they must smoothly connect to prevent discontinuities in the force experienced by atoms in the potential. This is done using sine functions that vary between 0 and 1 centered at a specified interatomic distance R and over a range of length D on either side of R. The smooth functions for potentials to the left and right of R are given below as Equations 2.6 and 2.7, respectively.

$$leftsmooth = \begin{cases} 1, & \text{if } r < R - D \\ \frac{1}{2} \times \left[1 - \sin\left(\frac{\pi}{2}\frac{r-R}{D}\right)\right], & \text{if } R - D < r < R + D \\ 0, & \text{if } r > R + D \end{cases} \tag{2.6}$$

$$rightsmooth = \begin{cases} 0, & \text{if } r < R - D \\ \frac{1}{2} \times \left[1 + \sin\left(\frac{\pi}{2}\frac{r-R}{D}\right)\right], & \text{if } R - D < r < R + D \\ 1, & \text{if } r > R + D \end{cases} \tag{2.7}$$

The distance at which to join the two potentials, R, and the distance over which to transition the functions on either side, D, are both parameters that can be adjusted. These smoothing functions are then multiplied by the original potential to create a function that smoothly falls off outside of its regime. An example of this method to join two potentials is illustrated in Figure 2.2. The 1D example features a short-range quadratic function with a long-range tail that falls off with the inverse of distance. In practice, SMRFF can be used to join Morse or Tersoff short-range potentials smoothly with LJ and Coulombic long-range interactions.

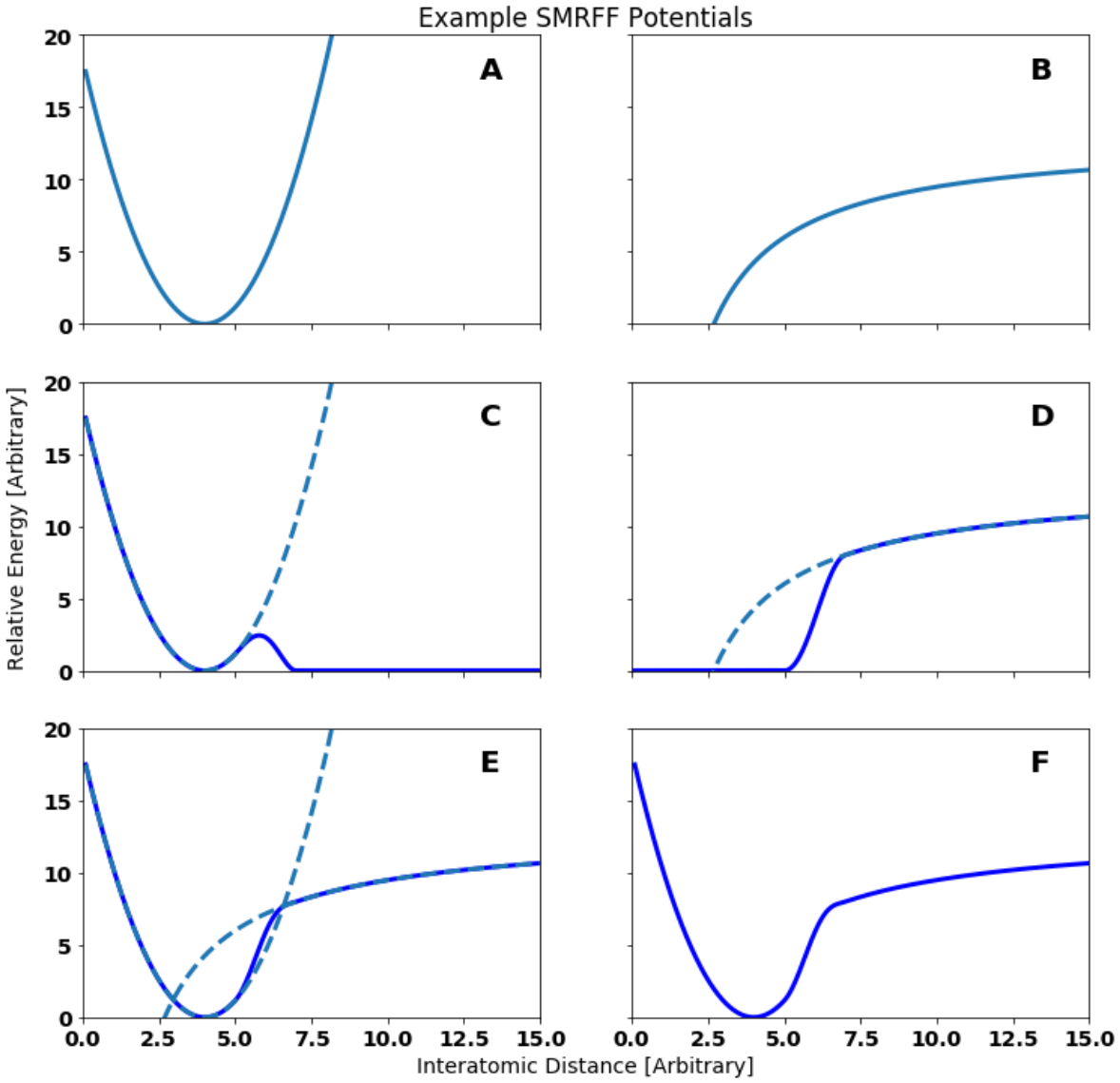


Figure 2.2: An example of how SMRFF joins potentials: A. a harmonic short-range potential; B. a decaying long-range potential; C. and D.; the potentials being smoothed to zero; and E. and F. the final SMRFF potential

The new forcefield was implemented into LAMMPS, an open-source MD solver from Sandia National Labs, to be used in molecular dynamics simulations [36]. In addition, we built a codebase that automates the parameterization process for the forcefield. Given target molecules to describe, the code will automatically handle everything from training set generation to parameter testing and optimization. Figure 2.3 depicts a flowchart of the internal process. First, the user specifies the target molecules and geometries to develop parameters for. The rest of the process is automated; first, the partial atomic charges on each atom are determined via DFT and optimally fit to maintain neutral molecules. Next, the seed molecules are perturbed to create the first training set. The code then calculates the DFT energies of the training set molecules. After collecting and cleaning the DFT data of non-converged calculations, the parameters are then optimized to fit the classical forcefield to the DFT data. Once the first round of parameters have been found, they are used in small MD simulations of the target molecules, from which the next iteration of seed geometries are taken. This process is continued for a set number of iterations until the final SMRFF parameters are achieved. Each step in the process is described in more detail in the following chapter.

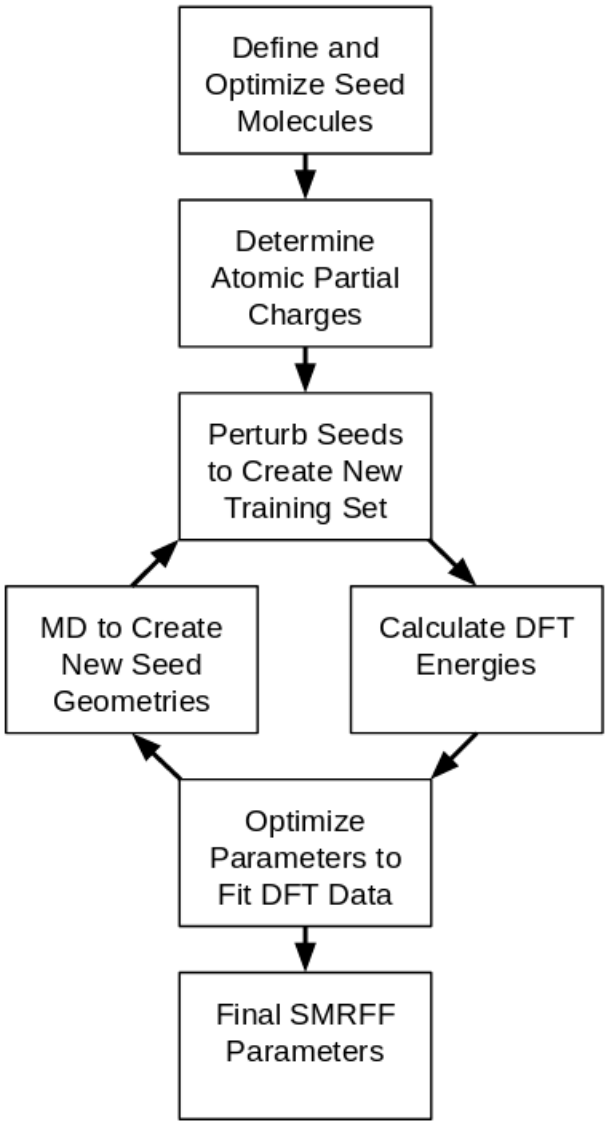


Figure 2.3: Flowchart of the SMRFF parameterization method

SMRFF Code Development

Developing a codebase that achieves automatic parameterization as described above involved the work of a small team. I detail my specific contributions to the project and how these methods operate below. This work is part of the in preparation manuscript *H. C. Herbol, G. Casee, W. L. Gao, O. Romilyui, J. Chaudhuri, and P. Clancy, The Simple Molecular Reactive Force Field Revisited - Extension to Directional Forces. (2019).* The codebase will become open source—along with supporting documentation so other research groups can use it in their own work—once the paper describing the method with accompanying case studies is published.

3.1 Partial Atomic Charge

Perhaps the most important parameters to determine accurately is the approximate partial atomic charge on each atom in a molecule since Coulombic forces dominate long-range interactions. Partial atomic charges are fixed to represent the electron distribution in a single molecule without the polarizing effect of neighboring species. Although polarizable forcefields are possible [37], they are computationally more expensive and difficult to parameterize. To approximate this effect, a Lennard-Jones potential is superimposed on the Coulombic potential. The Lennard-Jones potential is designed to capture the dispersion forces between (overall) neutral molecules. The attractive portion of the potential decreases with r^6, and the functional form of the repulsive portion has no physical significance. However, the SMRFF potential transitions to a more accurate short range description, effectively utilizing only the tail of the Lennard-Jones potential.

In this automated parameterization method, the charges are calculated and fixed first, separate from the rest of the potential parameters. Fixing atomic charges early reduces the challenge for the global optimizer later in the pipeline. Atomic charges cannot float freely to

any value as other parameters do since they are constrained to sum to the overall charge of the molecule. Global optimization problems, especially those consisting of dozens of parameters, are already difficult to solve without adding constraints on the parameters. In fact, many optimization schemes cannot handle constraints at all, aside from the required upper and lower bounds on the fitting parameters. Therefore, in order to simplify the optimization process and maintain flexibility in which optimization methods we choose to employ, it is beneficial to define atomic charges and optimize the other parameters around them.

In addition, partial atomic charges have a physical significance in that they represent the amount of electron density localized on a particular atom in a molecule. This is unlike most forcefield parameters, which often have nebulous underlying physical descriptions and are just used as fitting parameters to describe quantum behavior. Since atomic charges are quasi-physical, they can be calculated; DFT has several methods to estimate the partial atomic charges on atoms. I studied three major methods used to estimate charges from DFT results: population analysis, Charges from Electrostatic Potentials using a Grid-based method (CHELPG), and Atoms in Molecules (AIM).

There are several population analyses schemes, such as Mulliken [38] and Loewdin [39] studied here, but fundamentally they all estimate charge in the same way. Charges estimated via population analysis are calculated by assigning electrons to each atom in the system [40]. In quantum calculations, molecular orbitals are created by a linear combination of atomic orbitals, and the goal is to assign appropriate weights to these atomic orbitals to create an electron density. To estimate charge, electrons are then assigned according to these weights, and split evenly between overlapping orbitals when they are shared. The result is an estimate for how many electrons exist over each atom, which can then be subtracted from the number of protons in the atom to achieve a final partial charge. This method is the fastest to implement since it uses 'free' information that was already computed during the self-consistent field calculation.

CHELPG charges are calculated by creating a mesh grid around the molecule, calculating the electrostatic potential at each point in the mesh from the DFT output, and assigning charges to each atom to best fit the molecular electrostatic potential [41]. This method differs from population analysis schemes in that it does not allocate electrons to atoms, but rather assigns charges directly. CHELPG requires some post-processing of the DFT output by calculating the potential at all grid points from the electron density and performing an optimization to fit the atomic charges.

Finally, the AIM method is an extension of population analysis, but assigns electrons to atoms according to the electron density rather than to the weights of their atomic orbitals. The gradient of the electron density is examined to find critical points between atoms in which the density is at a minimum, and this point is taken as the demarcation between neighboring atoms. The amount of density over a particular atom relative to the entire molecule corresponds to a number of electrons that reside in that area, and the overall atomic charge is calculated. For this study in particular, I implemented the Bader method [42]. Like the CHELPG method, AIM requires post-analysis of the electron density to determine atomic charges.

I tested each method by estimating the partial charges of the atoms in a molecule of hydrogen cyanide (HCN). HCN was chosen for study because it is a small, polar molecule with three unique atoms, and thus unique charges. The molecule was optimized with the B3LYP functional and a def2-QZVP basis set. The resulting geometry was used for all further single point calculations. The functional was kept constant while the basis set was changed. The results are summarized below in Table 3.1.

Mulliken and Loewdin charges vary greatly with the change in basis sets and even swap signs of the charge in the case of C and N. In contrast, both CHELPG and Bader charges remain fairly consistent with the change in basis set. Additionally, both of those methods predict similar charges for the hydrogen atom; however, they vary greatly on the carbon and nitrogen

Table 3.1: Partial atomic charges for HCN calculated using four different methods with increasingly large basis sets

SVP	H	C	N
Mulliken	0.076	-0.082	0.006
Loewdin	0.038	-0.011	-0.027
CHELPG	0.202	0.120	-0.322
Bader (AIM)	0.232	0.829	-1.061
TZVP	H	C	N
Mulliken	0.203	-0.202	-0.001
Loewdin	0.125	-0.175	0.050
CHELPG	0.225	0.110	-0.334
Bader (AIM)	0.242	0.714	-0.957
QZVP	H	C	N
Mulliken	0.144	0.016	-0.161
Loewdin	-0.138	-0.037	0.175
CHELPG	0.221	0.116	-0.337
Bader (AIM)	0.238	0.726	-0.964

atoms. Bader assigns a far more negative charge to the nitrogen atom (and consequently, more positive charge to the carbon) than the CHELPG method does. As a result, Bader vastly over-estimates the dipole moment of the molecule at 6.50 debye, as compared to 2.99 from CHELPG, and 2.93 experimental [43].

Finally, the grid spacing was varied and the resulting CHELPG charges were calculated to see how this parameter affects accuracy. The results are summarized in Table 3.2. The charges are largely insensitive to changes in grid spacing for grids finer than the default 0.3 Angstroms. Smaller grid spacing requires more computational time since more points need to be evaluated and fit to charges.

In conclusion, the CHELPG method with a grid spacing of 0.3 Å is best for estimating the

Table 3.2: HCN CHELPG charges for increasing grid spacing, with smaller spacing corresponding to more sampled points and thus increased computational time.

Spacing [Å]	H	C	N
0.1	0.202	0.120	-0.322
0.2	0.201	0.122	-0.322
0.3	0.202	0.120	-0.322
0.4	0.201	0.122	-0.323
0.5	0.208	0.107	-0.315

atomic partial charges from DFT since it is independent of the chosen basis set and can recreate the experimental dipole moment of the target molecule.

Once the atomic charges are determined in DFT, they must be fit to each atom type that is being parameterized. The same atom type may appear multiple times in a given system, and each must have the same charge. In addition, the same atom type may also appear across entirely different molecules. For example, in a PbS nanocrystal, there are atoms on the surface as well as in the interior, and lead appears in the ligand as well. DFT will predict a different charge for each atom, but they should all have the same charge for simplicity in molecular dynamics. To solve this problem, I implemented a global optimizer to best fit the atom type charges to those predicted by DFT, with the constraint that the overall molecules should remain neutral (or have an integer charge, if an ion is specified). The method also collects charges that are already predetermined by OPLS (such as the hydrocarbon chains for the lead oleate ligand), then fits the other charges around them.

3.2 Training Set Generation

In order to fit any potential, data points are required to train the model. A training set consists of a group of molecular geometries along with their energies from DFT calculations. Having a large and varied training set is required for fitting parameters that are accurate over a range of conditions. However, creating training sets by hand can be time consuming

and not completely capture the range of possible conformations the molecule can explore. Therefore, the process of training set generation and expansion has been automated.

The first step to optimization is to generate training sets from the user-specified seed geometries. Given a location for the seed files, the method will collect those systems and create objects to store all geometries of the same type of system since DFT energies can only be compared across systems that contain the same number and type of atoms. For example, comparing the ground state energy of water to that of hexane is meaningless; only differences in energy between conformations of water (or hexane) are of any use. Therefore, the code was made to distinguish between systems and collect all that should be compared to each other during the error minimization.

Once the seed geometries are collected, they are perturbed to generate a training set. This is done in two major ways. First, the coordinates of each individual atom are perturbed by a random amount up to a maximum displacement distance. This creates training sets that are similar to the initial seed geometry and will train the forcefield to keep the molecule around this conformation. This is particularly useful for capturing the optimized geometry of the molecule and the most common conformations around it.

The second training set generation method is used to capture intermolecular interactions by generating systems with pairs or triplets of molecules. This method randomly perturbs each atom in a molecule as described above, then rotates each molecule randomly, and finally places the two molecules a set distance apart from each other. The process continues for increasing intermolecular distances. This method is used for parameterizing the long-range interactions between neighboring molecules and works best if intramolecular parameters have already been developed.

Finally, an extension of the routine to capture long-range interactions is used in solvent parameterization. The difference here is that a single solute of interest must be surrounded by many solvent molecules. I developed a method that attempts to place a solvent molecule

close to the solute, and then slowly tries larger distances until it can be placed such that it does not overlap with the solute. Solvent molecules are added in this way, each time attempting to insert as close to the solute as possible, and then rotating and pulling away until a suitable location is found.

Error handling is important to ensure the robustness of the training set generation routine; as such, the molecules are checked to ensure two atoms are not closer than a specified distance (default is 1.0 Å). Such overlaps will either result in unfeasibly high energies in DFT, or cause the self-consistent field calculations to fail entirely. Screening for possible atomic overlaps early saves computational time by not attempting lengthy DFT calculations that are destined for failure. Once the training set has been generated, the code calculates all of its DFT energies and collects the results for parameter fitting.

To use the training set generator, the user may specify the number of perturbations to make on each seed, the maximum displacement of each atom, the minimum approach of two atoms within the molecule, whether to create pairs and triples systems, as well as all relevant DFT parameters.

3.3 Training Set Expansion

Random perturbations of atomic coordinates can only generate training sets that are very similar to the seed geometries. Although one could (and traditionally would) create varied training sets by hand, this is time consuming, and it is difficult to create a comprehensive set of systems, especially as the molecules become more complicated. One approach to creating new training sets is to sample geometries from molecular dynamics. I implemented this method into SMRFF to automate the expansion of training sets so that the forcefield is trained to perform well under a variety of circumstances.

I implemented LAMMPS via its Python library interface commands to run MD simulations of the target molecules using the current parameters found by the optimizer. The seed geometries are placed into a large simulation box under an NVT ensemble and are allowed

to explore the space. Once the system has changed significantly from its initial state, the simulation is stopped and the coordinates of that system are used as a new seed for training set expansion. A criteria for when the system has changed significantly first had to be determined to know which frame to select for further expansion. This is quantified by calculating the change in the sum of all of the interatomic distances since the start of the simulation, as shown in Equation 3.1 below.

$$\Delta Geometry = \sum_{i,j}^{N} \frac{|r_{ij} - r_{ij_0}|}{r_{ij_0}} \tag{3.1}$$

Once this value exceeds a user defined amount that should be adjusted for each molecule, or if two atoms become unphysically close to each other, this frame is taken as a new expansion seed. If the geometry changed too much, the system is likely in an unfavorable position on the energy landscape. However, the still-naive classical forcefield predicts that this state is still accessible at a given temperature. Once this new geometry, as well as perturbations about it, are added to the training set, the re-parameterized forcefield will predict that that area is high energy, and future MD simulations will follow a different trajectory.

Training set expansion via molecular dynamics is an efficient method to generate varied molecular geometries that provide new information for parameterization. Selecting new geometries in this way biases the training towards areas in which the forcefield is not currently working. With added parameterization cycles, the forcefield learns from its past failures and becomes better at realistically describing the target molecules.

3.4 Parameter Optimization

Minimizing an error function by changing possibly dozens of parameters presents a difficult optimization problem to solve efficiently. To handle this, the open source library NLopt (nonlinear optimization) was implemented into the SMRFF codebase to handle parameter optimization [44]. Given an objective function and a list of parameters, NLopt efficiently

implements a user-specified algorithm to optimize the given parameters. The objective function to be minimized is the same as in my previous PbS parameterization described above (Equation 2.5). This error is a modified sum of least squares by dividing the square error by the DFT energy of the data point to prevent high energy systems from skewing the potential fitting. To minimize this function, the optimization scheme first performs a global optimization on the parameters, followed by a local optimization to refine the results and fall into the local basin.

Since it is impossible to be sure than a global optimizer finds the global minimum of the error function without enumerating all possible values for each parameter, it is not initially clear when to stop the global optimization routine. In fact, global optimization will continue to decrease the error over time, albeit at a slower rate. The optimization can either be given a hard time limit (either through specifying a maximum wall time or number of function calls) or a tolerance criteria to stop if the objective function only decreases by a small amount on each iteration. In my experience testing this code, however, tolerance criteria do not stop a global optimizer in a desirable way. Oftentimes, the optimizer will be 'stuck' at a certain error level before finally exploring a new region of parameter space and experiencing a sharp decline in total error. If a tolerance is used to stop the optimizer, it will often end prematurely before experiencing a sharp improvement. Therefore, setting a time limit on optimization ensures this does not happen and is the default behavior of the routine.

The method of molecular dynamics expansion generates an increasingly large training set upon each iteration, which results in a decrease in the number of function calls the optimizer can perform in a given time limit. This undesirable behavior leads to parameters that actually perform worse as the parameterization progresses! To remedy this, the time for global optimization is scaled accordingly such that the program achieves approximately the same number of function calls on each iteration. This ensures that the relative accuracy for the parameters is the same for each iteration. Otherwise, the molecular dynamics simulation

uses non-optimal parameters and consequently fails in uninformative ways.

Using the optimizer is made to be simple; no input is required, although the user may specify and global and local optimization methods to use, the time to allow the optimizers to run, and whether to increase the time spent on each optimization on subsequent iterations. The results are passed on to the molecular dynamics expansion code to generate more training sets and continue the cycle.

Parameterizing SMRFF for Quantum Dot Systems

This section presents two examples of how the paramerization method detailed above can be used to generate parameters for PbS quantum dots as well as their passivating ligands. The methods in which the crystalline quantum dot and the ligands are parameterized are necessarily different. The former must capture bulk behavior along with surface effects, while the latter is must describe a complete and distinct molecule. Therefore, the types of seeds that are used to train the models, as well as the way in which training sets are created, differ for each system. The methods used to parameterize these two systems, as well as the parameterization results, are detailed below. In both cases, the SMRFF potential consists of a short-range Morse potential that smooths to long-range Lennard-Jones and Coloumbic forces.

4.1 Parameterizing Lead-Sulfide Nanocrystals

Since studying an experimentally relevant-sized quantum dot is not possible due to the large number of atoms, smaller seed molecules need to be carefully selected that will recreate the dynamics expected for a large nanocrystal. However, as the number of atoms in the seed begin to exceed 50 to 60, the DFT calculations start to become prohibitively long due to the massive number of electrons contained in each lead (84)—and to a lesser extent, sulfur (16)—atom. With about 60 atoms as the upper limit for seed size, there are a number of geometries that can be conceived to train the forcefield. Since we are interested in the interaction of lead oleate ligands on nanocrystal surfaces, a seemingly natural choice might be to create seeds that recreate the three main facets on a PbS nanocrystal. Examples of the training seeds are shown in the top half of Figure 4.1. However, as shown and discussed below, using these seeds results in nonphysical quantum dot dynamics. 56 atom mini quantum dots were used instead

as seeds for the final nanocrystal parameterization. Perfect FCC structures were created with lattice constants varying from 5.6 Å to 6.3 Å to capture the energy landscape around the experimental lattice constant of 5.936 Å [45]. Examples of these seeds are depicted in the lower half of Figure 4.1. Finally, I note that creating training geometries of randomly placed atoms fails spectacularly, despite sampling perhaps the largest region of the potential energy landscape and having the most varied geometries.

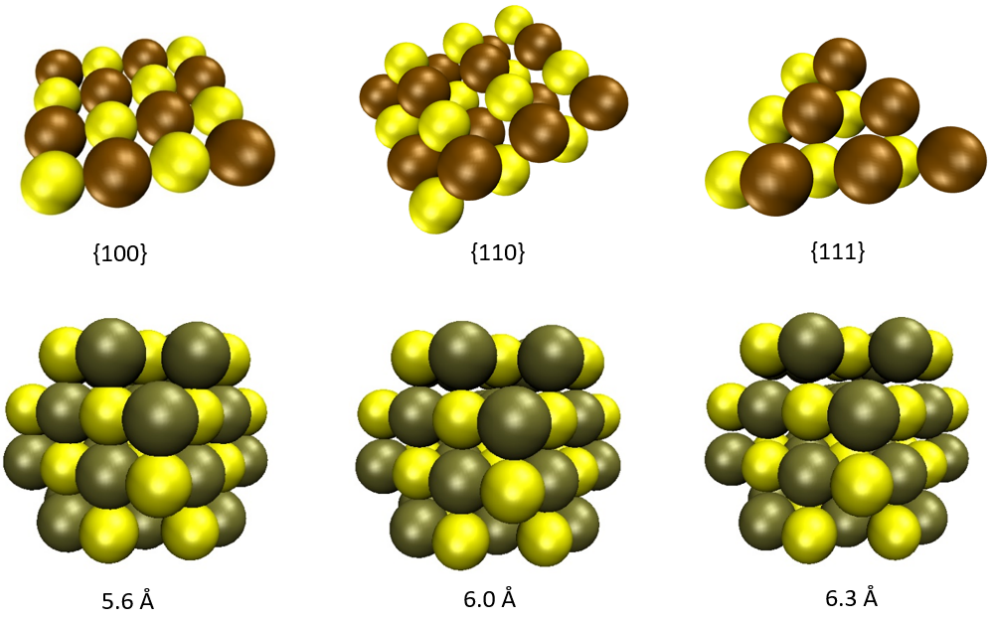


Figure 4.1: Seed geometries used to train PbS quantum dots. The lower geometries were used to generate the final parameters while the upper result in erroneous dynamics.

Figure 4.2 highlights how the choice of training geometries affects the resulting dynamics of the system. The first dot was largely trained with geometries representing 100, 110, and 111 surfaces, while the second was trained using mini-dots of 56 atoms. Although 32 atom mini-dots are the smallest representation that contains (however small) fully coordinated interior atoms, 100, 110, and 111 facets, slightly larger structures are needed to properly maintain crystallinity in the larger quantum dot simulations.

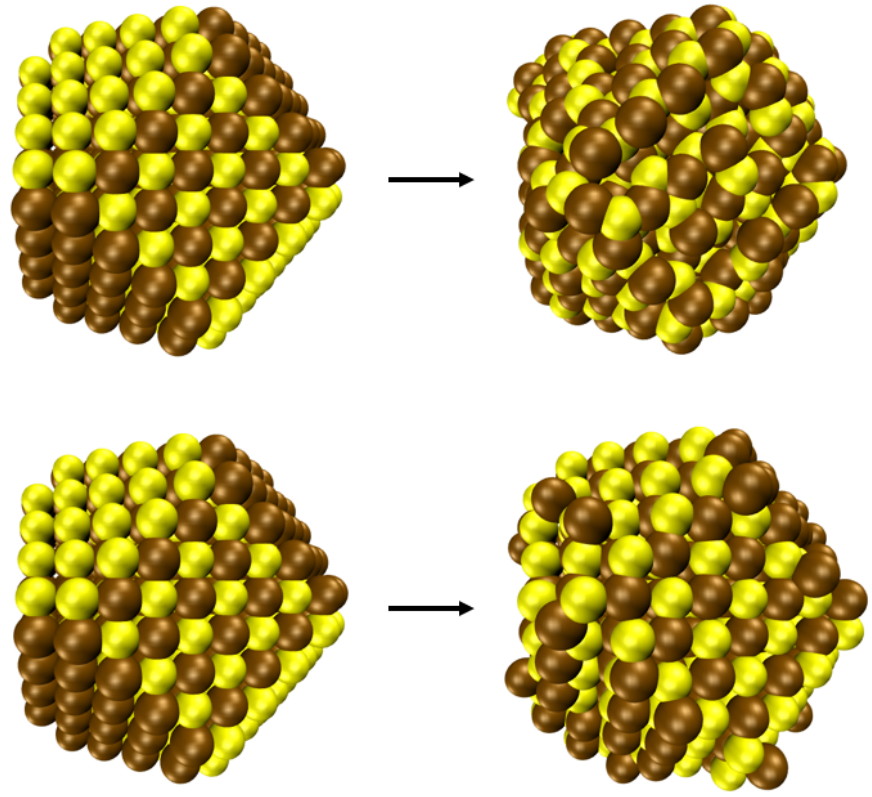


Figure 4.2: Molecular Dyanmics results of 720 atom PbS dots that were parameterized with surface geometries (above) and crystalline geometries (below).

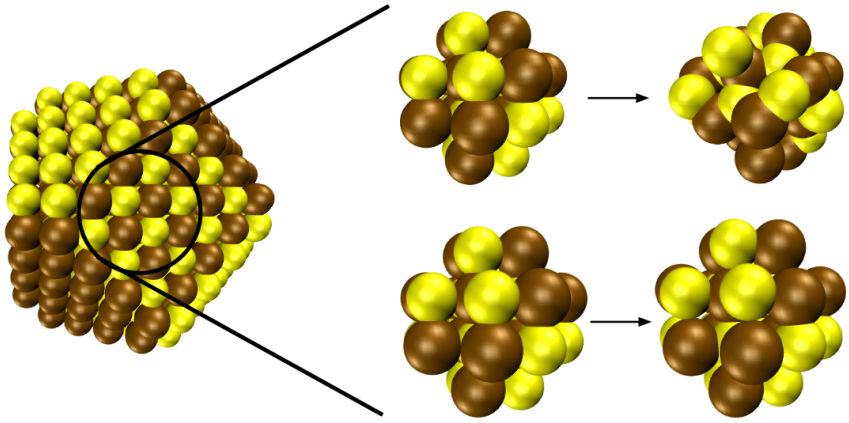


Figure 4.3: Evolution of the core atoms in the two quantum dots shown above; the atoms of the dot trained on surfaces (above) become amorphous as compared to the crystalline core of the second dot (below).

As shown in Figures 4.2 and 4.3, training the forcefield on 56 atom 'mini-quantum dots' results in a quantum dot that largely maintains its structure. To confirm that the dot dynamics are realistic, a radial distribution function was plotted to watch how the positions of the atoms change over time. Figure 4.4 shows the plot for the final frame of the MD simulation depicted in Figure 4.2.

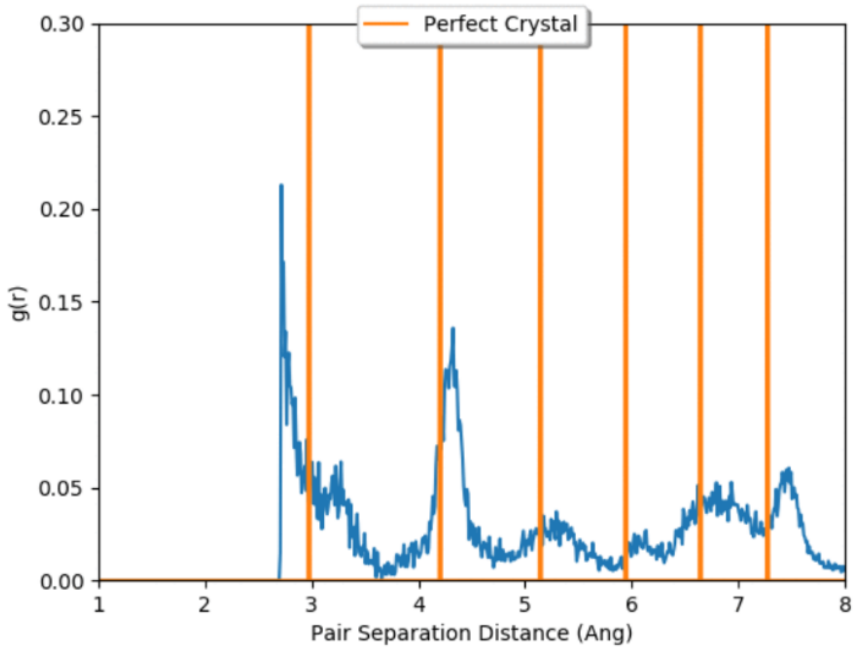


Figure 4.4: Radial distribution function for 720 PbS dot after relaxation in molecular dynamics. The vertical lines represent the distribution for a perfect, bulk PbS crystal.

The peaks of the radial distribution function roughly align with those expected in bulk PbS crystal. The first peak is under-predicted; however, this is to be expected as nanocrystals will slightly contract to minimize the exposed surface. The rest of the peaks align well with what is expected in a perfect crystal, with noise that is expected from the randomness of MD as well as the non-crystalline structure of the surface atoms.

It is expected that quantum dots rearrange and deviate from the initial perfectly octahedron shape to passivate lead and sulfur terminated surfaces. In my simulations, however, it is always the lead atoms that can be seen diffusing across the crystalline surface and embedding

themselves in the sulfur terminated 111 facets. To explore why this may be the case, I looked at pairwise interactions for lead and sulfur atoms and determined Morse parameters for each, as summarized in Table 4.1. The dissociation energy for lead is less than that of sulfur, suggesting that it takes less energy for it to move away from its initial crystalline position. Additionally, the equilibrium distance for lead interactions is less than that of sulfur bonds. These two points suggest that it may be easier for lead to mobilize on the quantum dot surface, as seen in my simulation results.

Table 4.1: Morse parameters fit to pairwise energies between Pb and S

Pair	D_0 [kcal/mol]	alpha [$Å^{-1}$]	r_0 [Å]
Pb-Pb	54	1.56	2.61
Pb-S	119	1.41	2.29
S-S	146	1.51	1.92

4.2 Parameterizing Lead-Oleate Ligands

The passivating agents that bind to the surface of colloidal quantum dots are often lead oleates. Oleate is an 18-carbon chain ($C_{18}H_{33}O_2^-$) and two chains bind to each lead atom. Therefore, a single lead oleate ligand contains 107 atoms, and—more importantly to DFT—396 electrons. Although performing DFT calculations on entire lead oleate ligands is easily attainable, the calculation times become cumbersome once we begin expanding the training set, or looking at interactions between ligands and crystalline PbS structures. The chains beyond the carboxylate head group are already well described by the OPLS forcefield, so any perturbations to these hydrocarbon tails largely waste computational time without adding relevant information to the parameters of interest. Therefore, it is desirable to parameterize shorter carboxylate chains while maintaining the electronic structure of the head group.

As in the section on binding energy calculations, the effects of ligand length are negligible even at only 2 carbon atoms when simply considering binding energies in DFT. Although

the binding energy is largely independent of ligand length, the length of the ligand affects the electronic structure around the head group and may influence how it interacts with its neighbors in non-minimized geometries. I determined how long the hydrocarbon chain must be by using the CHELPG method to estimate partial atomic charges and using them as a proxy for electronic structure. First, the atomic charges for the lead, oxygen, and first two carbon atoms were determined for the full lead oleate ligand. This calculation was repeated for optimized lead carboxylate structures up to pentanoate, and the charge results are listed in Table 4.2 below along with percent differences from the charges in the full oleate chain.

Table 4.2: Partial atomic charges for the first four atom types in lead carboxylate ligands of increasingly length compared to the full oleate chain.

Chain Length	Pb charge	O charge	C1 charge	C2 charge
Oleate	0.615	-0.464	0.690	-0.453
C2	0.615 (0.0%)	-0.454 (-2.1%)	0.714 (3.6%)	-0.700 (54.6%)
C3	0.628 (2.0%)	-0.450 (-3.1%)	0.577 (-16.3%)	-0.071 (-84.2%)
C4	0.624 (1.4%)	-0.459 (-1.1%)	0.648 (-6.1%)	-0.434 (-4.3%)
C5	0.609 (-1.0%)	-0.468 (0.9%)	0.736 (6.7%)	-0.294 (27.0%)

A carboxylate chain containing four carbons is sufficient to recreate the partial charges of the first few atoms on a full oleate tail. We can drastically expedite the DFT calculations needed to create training data for parameterization by shortening the ligand from 107 atoms and 396 electrons to only 27 atoms and 176 electrons.

To maintain compatibility with the PbS forcefield, the ligand was parameterized to utilize the Morse short-range potential with long-range smoothing to Lennard-Jones and Coulomb. However, only the head group acts according to Morse; the hydrocarbon chain utilizes existing OPLS parameters. The oxygen atoms are hybrids that link the two regimes; they are described by Morse when interacting with lead and other oxygen atoms, and are described by OPLS when interacting with the rest of the chain.

Unlike nanocrystals, the ligands have a well defined size and structure to be captured by the forcefield. Therefore, a single, energy minimized seed of lead butanoate was used as the starting point for parameterization and I relied heavily on MD expansion to generate training sets. Molecular dynamics captured many unique conformations of the chain as well as positions of the two carboxylates around the central lead atom, and ultimately led to a varied training set after many iterations. The parameterization was considered successful once the molecule could bounce around in molecular dynamics indefinitely without falling apart or imploding.

In conclusion, the automated parameterization method successfully fits parameters to a SMRFF forcefield that joins Morse and LJ/Coulombic potentials. Both PbS quantum dots and their passivating ligands can be described by this potential, thus granting the ability to study these systems on large scales.

Improvements to Forcefield Parameter Fitting

The methods proposed in this final chapter are designed to improve forcefield performance without generating additional training data. In the SMRFF scheme proposed above, they could be implemented in the parameter optimization step when evaluating the error function. However, these techniques are general and independent of the atomic potential that is being fit. Therefore, they can—and arguably should—be applied to any forcefield parameterization problem.

5.1 Similarity Correction for Training Sets

One issue that may arise while creating training sets is that many geometries may be similar to each other and thus provide little new information to forcefield training. In fact, training sets that are similar will unnaturally pull any regression towards them, skewing the resulting potential. A simple example to highlight the issue is shown below in Figure 5.1. In this example, extra points are added around $x = 8$ to illustrate how they pull the regression to fit the extra data.

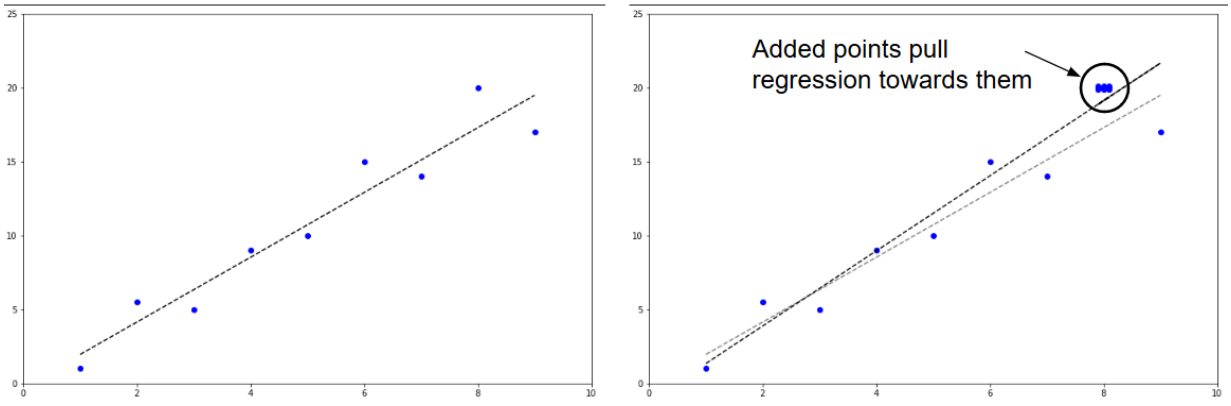


Figure 5.1: The addition of extra data points in close proximity skew regressions to fit the redundant data.

Although harder to detect, the same phenomena can happen when fitting complex functions to three-dimensional data instead of this linear example. If training geometries are created that are very similar to each other, it is purely incidental and a result of the training set generation method. That is, there is no physical justification for why one point (and points around it) in configurational space should be represented numerous times. Traditional error calculation methods would effectively assign more weight to areas with more data even if that region is not likely to be explored more than others.

To correct for this, I developed a function that will determine the similarity between two interatomic distances and assign weights to the training points accordingly. The function is given below as Equation 5.1 and has the properties that it outputs values between 0 and 1, the similarity of any data point with itself is 1, and the similarity of the most extreme data points (x_L and x_R) is defined as 0.

$$s_{ij} = 1 - \left(\frac{x_i - x_j}{x_R - x_L} \right)^2 \tag{5.1}$$

The similarity indices are then turned into a weight for the data point by summing over all similarities and inverting the sum, as shown in Equation 5.2. Note that the data point itself is included in the sum; this ensures that the minimum value the sum takes is 1 so the inverse is well-behaved. Weighting the data in this manner assigns less importance to values that are more similar to the rest of the data set.

$$W_i = \left(\sum_j^N s_{ij} \right)^{-1} \tag{5.2}$$

5.2 Boltzmann Weighting for Parameter Fitting

As discussed earlier in detailing the form of the error function for fitting forcefield parameters, simply minimizing the sum of the square differences between classical and DFT energy differences may not be the most accurate approach since unfeasibly high energy systems will pull the regression towards them. As shown before, one correction is to divide the energy difference by the energy of the system. That is, high energy states are divided by comparatively large numbers and thus contribute less to the error function. Although this addresses the issue, it is not the best solution from a thermodynamic standpoint.

The naive approach of minimizing the sum of the square differences between all classical and DFT energies implies that each system is of equal importance. Thermodynamically, this is only true at infinite temperature, in which all states of a system are equally accessible. For all other temperatures, the Boltzmann distribution describes the frequency distribution of states in a system. In particular, the Boltzmann factor defines the relative probabilities of observing states i and j, as reproduced below:

$$\frac{P_i}{P_j} = e^{\frac{-(E_i - E_j)}{k_b T}} \tag{5.3}$$

If we set E_j to be our minimum energy in the DFT training set, we get an expression for the relative weights of each energy:

$$W_i = e^{\frac{E_{min} - E_i}{k_b T}} \tag{5.4}$$

in which W_i is the weight of data point i, and the weight of the minimum energy system will be exactly 1. Note that when written in this form, all the weights fall between 0 and 1. Any arbitrary reference point could instead be taken which would result in a different range of weights, but with the same ratios between any two given points. Therefore, the choice is

arbitrary, but setting the reference to the minimum energy provides a clearer picture of which points are most important. This resulting list of weights defines the likelihood of observing state i relative to the minimum energy state if the system was free to explore space at a given temperature, T. Figure 5.2 illustrates the effects of Boltzmann weighting on fitting a Morse potential to diatomic nitrogen at a temperature of 298K.

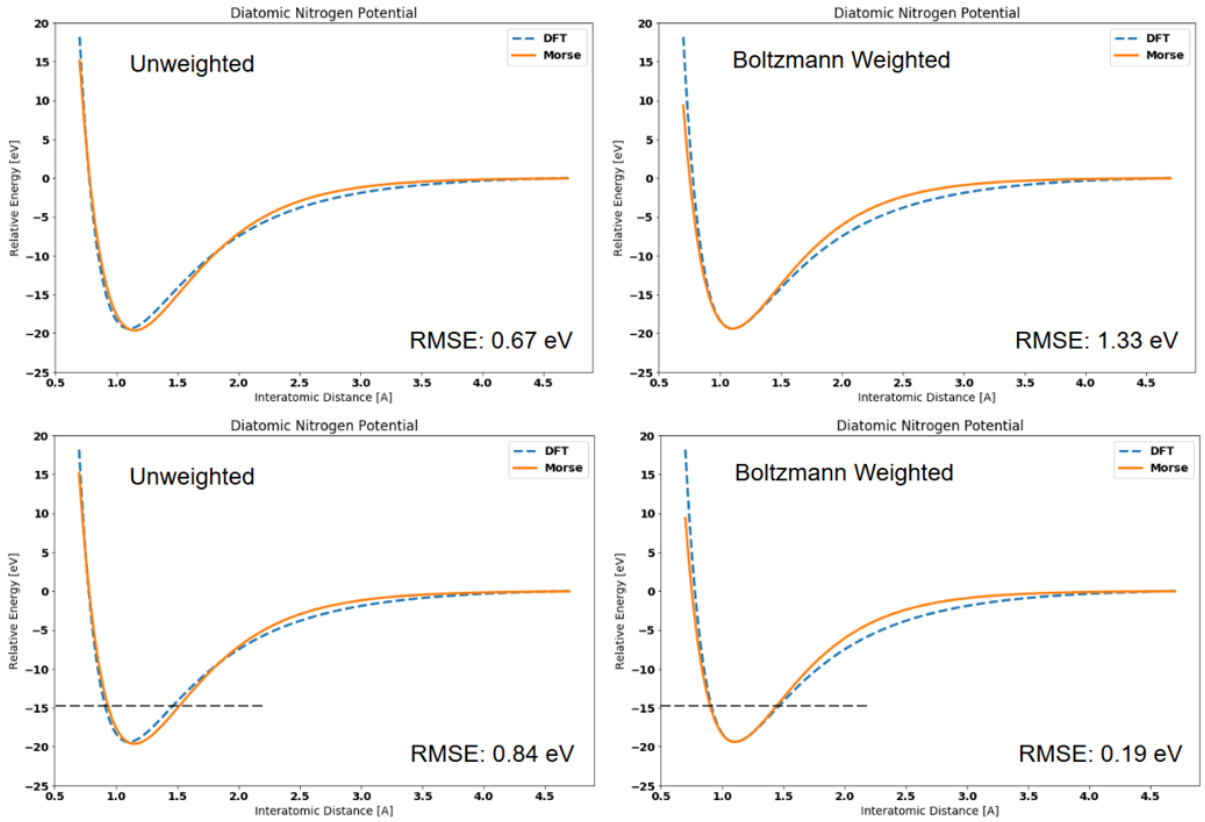
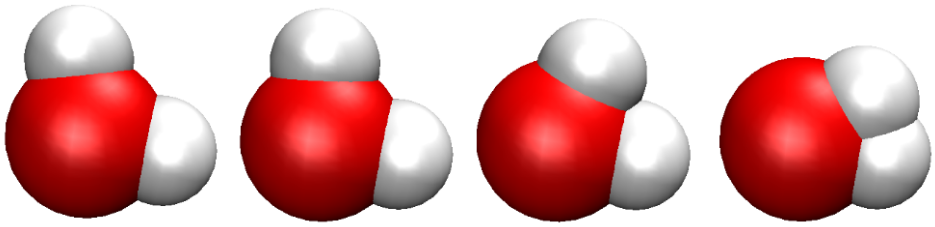


Figure 5.2: Morse potential fit to diatomic nitrogen using least squares (left) and least squares with Boltzmann weighting (right) regressions. Although the overall RMSE increases when the weighting is applied (upper plots) the error decreases in the energy basin (lower plots).

Although the overall error while fitting to DFT energies is larger (RMSE of 0.67 eV versus 1.33 eV), this is not the error that should be minimized. For example, in a molecular dynamics simulation of a single molecule, the error that should be minimized is the difference in the classical force field energy and the DFT energy for all the explored conformations of the molecule. If the temperature is low enough, most of the high energy states will not be

explored at all, and it is irrelevant how well the forcefield fits at those conditions. To illustrate the benefit of Boltzmann weighting, the error is recalculated only for data points within 5 eV of the energy minimum. Here, the RMSE of drops from 0.84 eV in the unweighted case to 0.19 eV in the weighted example.

Molecules in real systems, however, are subjected to energies far higher than thermal energy. For example, the energy of a hydrogen bond can be as high as 0.3 eV—over 12 times thermal energy at room temperature [46]. Thus, the probability of a hydrogen bond in water, if calculated using Equation 5.3 at room temperature, approaches zero. This is clearly a nonphysical result, as liquid water is known to be constantly making and breaking such bonds [47]. If instead the water is weighted by the energy of a hydrogen bond which it is reasonably expected to be subjected to, the probability for previous inaccessible states becomes appreciable. An example to illustrate this is shown below in Figure 5.3



Thermal Energy Reference (298K)	1.00	0.15	0.00	0.00
Hydrogen Bond Energy Reference	1.00	0.85	0.06	0.00

Figure 5.3: Boltzmann weights for different configurations of a water molecule calculated relative to thermal energy at room temperature and the energy of a hydrogen bond.

Boltzmann weighting provides a natural and physically significant choice of weighting since it represents the relative probability of finding the system in one state as compared to another at a given energy. Fitting without such weighting is equivalent to assuming each data point is equally probable, which is only true at infinite temperature. Barring this extreme case, all parameterizations should be done with a target energy scale to accurately describe the

conformations of physical significance.

5.3 Error Comparison for Weighting Methods

The methods proposed in the prior sections are intended to create a training set whose constituent molecules and energies are most similar to what would be found in nature. Since it is impossible to sample geometries and energies from a real system (that is, take atomistic snapshots from real molecules in a flask), the existing data is weighted to give the illusion that those points are sampled more or less frequently to approach what is thermodynamically expected. This is tested by fitting a forcefield with both weighting methods, as well as a combination of the two. In order to create a training set that is perfectly drawn from the Boltzmann distribution, all points need to be equally represented. For example, if a particular training set were to be duplicated, its effective weight would be twice that of its Boltzmann weighting. If two training geometries are similar but not exact, they still contribute more effective weight to a particular region in space. Therefore, the Boltzmann weighting is best combined with a similarity weighting scheme to unbias redundant training data. A method to simulate sampling points from a molecular dynamics simulation was employed to evaluate the efficacy of these weighting methods, as detailed below.

First, a random set of interatomic distances for nitrogen are selected. In this case study, 10, 20, or 40 data points are chosen. The Morse potential is then fit to these data points and their respective energies. The fitting is done in one of four ways: first, by simply minimizing the sum of least squares; second, by weighting the data points by a similarity factor to unbias regions with a high density of points; third, by minimizing the sum of least squares weighted by a Boltzmann factor that is calculated based on the system's energy; and fourth, by combining the Boltzmann weight with the similarity correction.

The performances of the resulting potentials are calculated in two different ways. The first is to find the RMSE when considering all data points in the range from 0.7 to 4.7 Å once. This represents the typical method for evaluating a forcefield's fit. The second is through a Monte

Carlo simulation to mimic how the forcefield would perform in a real MD simulation. In the second method, nitrogen configurations are selected randomly from a Boltzmann distribution calculated from the DFT data. This is analogous to sampling configurations from an actual molecule of nitrogen if it were isolated. The probability density function for the range of nitrogen interatomic distances considered is shown in Figure 5.4. The temperature was increased to 50,000 K so that all states are reasonably accessible. To decide how many points to sample, I took the inverse of the smallest weight and rounded up. According to the binomial distribution, this ensures that, on average, every point is sampled at least once. For this system, approximately 170,000 steps are required for it to be likely that every state is visited.

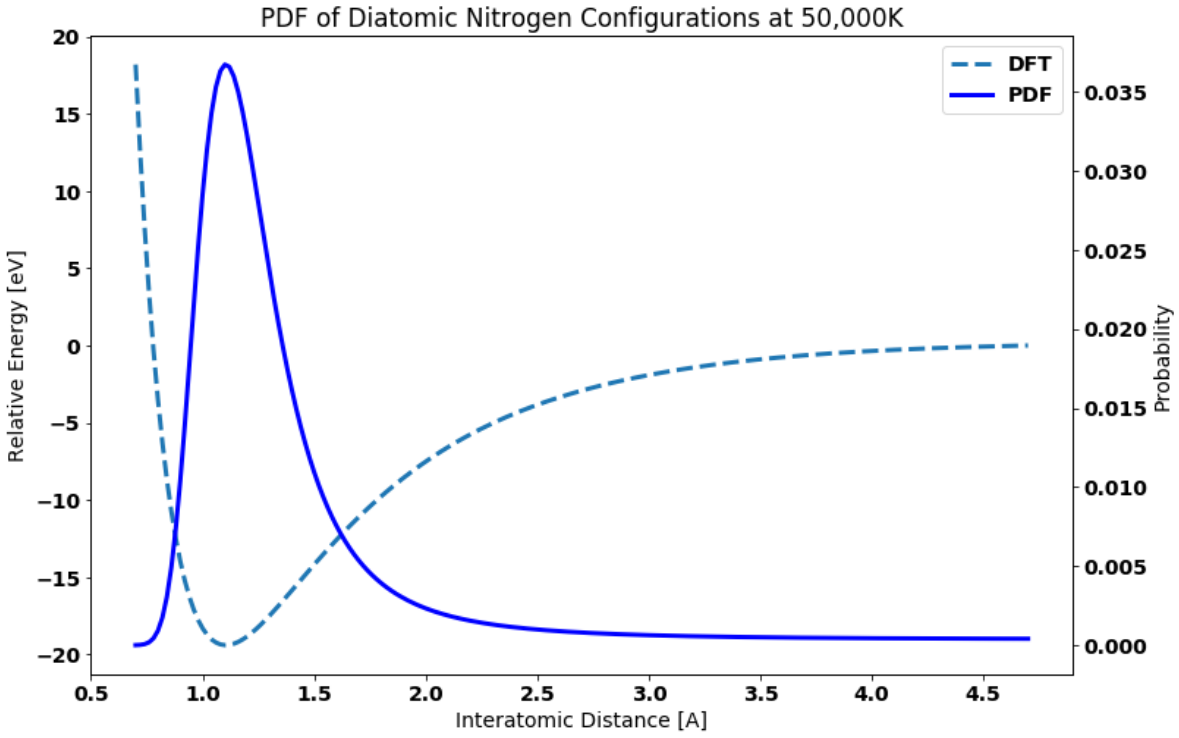


Figure 5.4: Probability density function of diatomic nitrogen at 50,000K superimposed on the DFT data.

Once errors are calculated for a given set of training data, this entire process is repeated with a new set of random points to generate statistics for the mean error expected when fitting a potential. However, the calculated errors are always skewed right and thus require some

post-processing to get meaningful statistics. The skewness is expected since the minimum error that can be obtained is zero (a perfect fit), while the maximum error is unbounded. To normalize the data, I implemented the one-parameter Box-Cox transformation [48]. The optimal transformation parameter, λ, was determined for each data set by maximizing the correlation between the transformed data and normally distributed z-scores in a QQ plot. In general, the data is best made normal for λ values of around -0.5. Once the data is normalized, the mean and standard error can be calculated and transformed back into the original units for comparison. The results are detailed in Table 5.1.

Table 5.1: Average errors between DFT data and a Morse potential fit with an increasing number of points and various weighting methods. The top results are for fitting all equally spaced data points, and the bottom results are from the MC simulation. All results are reported as Mean ± Standard Error in eV.

All data RMSE				
Training Points (n trials)	Unweighted	Similarity Weighted	Boltzmann Weighted	Sim and Boltz
10 (n=569)	2.074 ± 0.066	2.051 ± 0.065	1.965 ± 0.058	1.939 ± 0.057
20 (n=560)	1.444 ± 0.038	1.425 ± 0.037	1.397 ± 0.026	1.351 ± 0.024
40 (n=561)	1.196 ± 0.023	1.199 ± 0.023	1.418 ± 0.021	1.391 ± 0.021
Monte Carlo RMSE				
Training Points (n trials)	Unweighted	Similarity Weighted	Boltzmann Weighted	Sim and Boltz
10 (n=569)	1.391 ± 0.042	1.361 ± 0.041	1.151 ± 0.039	1.144 ± 0.039
20 (n=560)	0.976 ± 0.021	0.954 ± 0.021	0.642 ± 0.014	0.635 ± 0.014
40 (n=561)	0.911 ± 0.018	0.902 ± 0.018	0.772 ± 0.022	0.776 ± 0.022

A few trends are immediately noticeable from the results. First, the errors decrease across the board as the model is trained on more points. This is unsurprising as more data should improve forcefield performance, but it still serves as verification that the model is working as intended.

The similarity weighting alone only provides a slight improvement over the unweighted case. Furthermore, the improvement appears to decrease as the potential is trained on more data points. This may be due to the fact that the points are chosen randomly from a uniform distribution. With more sampled points, data tend not to be clustered in particular regions of space, and thus there is no bias for the similarity weighting to correct for.

The Boltzmann weighting delivers the most significant improvement to forcefield performance when evaluating the error according to the Monte Carlo method described above. This is due to the phenomena depicted in Figure 5.2 in which the forcefield is biased to fit better at lower energy states at the expense of not fitting higher energy systems as well.

Finally, combining the similarity weighting with the Boltzmann weighting appears to slightly improve the performance. This may indicate that the two methods are indeed correcting for two different problems, and can be used in conjunction to get the best fit with the same data points.

In conclusion, Boltzmann weighting greatly improves forcefield performance by better describing molecules in the states that are more likely to be explored in MD, while similarity weighting corrects for having high density of points in the training set. Together, they provide a computationally inexpensive way to produce better performing molecular forcefields without requiring additional—and potentially costly—training data.

References

(1) Ekimov, A. I.; Onushchehko, A. A. Quantum Size Effect in Three-Dimensional Microscopic Semiconductor Crystals. *ZhETF Pis ma Redaktsiiu* **1981**, *34*, 363.

(2) Ashoori, R. C. Electrons in artifical atoms. *Nature* **1996**, *379*, 413–419.

(3) Whitham, K. et al. Charge transport and localization in atomically coherent quantum dot solids. *Nature Materials* **2016**, *15*, 557–563.

(4) Evers, W. H. et al. Low-Dimensional Semiconductor Superlattices Formed by Geometric Control over Nanocrystal Attachment. *Nano Letters* **2013**, *13*, 2317–2323.

(5) Nie, Z.; Petukhova, A.; Kumacheva, E. Properties and emerging applications of self-assembled structures made from inorganic nanoparticles. *Nature nanotechnology* **2010**, *5*, 15–25.

(6) Kalesaki, E. et al. Electronic structure of atomically coherent square semiconductor superlattices with dimensionality below two. *Physical Review B* **2013**, *88*.

(7) Semonin, O. E.; Luther, J. M.; Beard, M. C. Quantum dots for next-generation photovoltaics. *Materials Today* **2012**, *15*, 508–515.

(8) Feng, J. et al. Single-layer MoS2 nanopores as nanopower generators. *Nature* **2016**, *536*, 197–200.

(9) Baumgardner, W. J.; Whitham, K.; Hanrath, T. Confined-but-Connected Quantum Solids via Controlled Ligand Displacement. *Nano Letters* **2013**, *13*, 3225–3231.

(10) Whitham, K.; Smilgies, D.-M.; Hanrath, T. Entropic, Enthalpic, and Kinetic Aspects of Interfacial Nanocrystal Superlattice Assembly and Attachment. *Chemistry of Materials* **2018**, *30*, 54–63.

(11) Drijvers, E. et al. Ligand Displacement Exposes Binding Site Heterogeneity on CdSe Nanocrystal Surfaces. *Chemistry of Materials* **2018**, *30*, 1178–1186.

(12) Geuchies, J. J. et al. In situ study of the formation mechanism of two-dimensional superlattices from PbSe nanocrystals. *Nature Materials* **2016**, *15*, 1248–1254.

(13) Brown, P. R. et al. Energy Level Modification in Lead Sulfide Quantum Dot Thin Films through Ligand Exchange. *ACS Nano* **2014**, *8*, 5863–5872.

(14) Dubois, F. et al. A Versatile Strategy for Quantum Dot Ligand Exchange. *Journal of the American Chemical Society* **2007**, *129*, 482–483.

(15) Kaushik, A.; Clancy, P. Solvent-Driven Symmetry of Self-Assembled Nanocrystal Superlattices–A Computational Study. *Journal of Computational Chemistry* **2012**, *34*, 523–532.

(16) Weidman, M. C. et al. Impact of Size Dispersity, Ligand Coverage, and Ligand Length on the Structure of PbS Nanocrystal Superlattices. *Chemistry of Materials* **2018**, *30*, 807–816.

(17) P. Perdew, J.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical review letters* **1996**, *77*, 3865–3868.

(18) Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter* **2009**, *21*, 395502.

(19) Blochl, P. Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Phys. Rev.* **1994**, *50*.

(20) Kresse, G.; Joubert, J. From ultrasoft pseudopotentials to the projector augmented wave method. *Phys. Rev.* **1999**, *59*.

(21) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.

(22) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**.

(23) Klamt, A.; J. Schueuermann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans. 2* **1993**, *2*, 799–805.

(24) R Bealing, C. et al. Predicting Nanocrystal Shape through Consideration of Surface-Ligand Interactions. *ACS nano* **2012**, *6*, 2118–27.

(25) Stevenson, J.; Ruttinger, A. W.; Clancy, P. Uncovering the reaction mechanism initiating the nucleation of lead sulfide quantum dots in a hines synthesis. *J. Mater. Chem. A* **2018**, *6*, 9402–9410.

(26) Levitt, M. Molecular dynamics of native protein: I. Computer simulation of trajectories. *Journal of Molecular Biology* **1983**, *168*, 595–617.

(27) Berendsen, H.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995**, *91*, 43–56.

(28) Jones, J. On the determination of molecular fields. -II. From the equation of state of a gas. *Royal Society* **1924**.

(29) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988**, *110*, PMID: 27557051, 1657–1666.

(30) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

(31) Morse, P. M. Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Phys. Rev.* **1929**, *34*, 57–64.

(32) Tersoff, J. New empirical approach for the structure and energy of covalent systems. *Phys. Rev. B* **1988**, *37*, 6991–7000.

(33) Schapotschnikow, P. et al. Morphological Transformations and Fusion of PbSe Nanocrystals Studied Using Atomistic Simulations. *Nano Letters* **2010**, *10*, 3966–3971.

(34) Zherebetskyy, D. et al. Hydroxylation of the surface of PbS nanocrystals passivated with oleic acid. *Science* **2014**, *344*, 1380–1384.

(35) Andrejevic, J.; Stevenson, J.; Clancy, P. Simple Molecular Reactive Force Field for Metal-Organic Synthesis. *Journal of Chemical Theory and Computation* **2016**, *12*, 825–838.

(36) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Chemical Physics* **1995**, *117*, 1–19.

(37) Warshel, A.; Kato, M.; Pisliakov, A. Polarizable Force Fields: History, Test Cases, and Prospects. *Journal of Chemical Theory and Computation* **2007**, *3*, PMID: 26636199, 2034–2045.

(38) Mulliken, R. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. *Journal of Chemical Physics* **2004**, *23*.

(39) Loewdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *Journal of Chemical Physics* **1950**.

(40) Hunt, P. Molecular Orbitals and Population Analysis. **2008**.

(41) Breneman, C.; Wiberg, K. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry* **1990**.

(42) Bader, R. F. W. A quantum theory of molecular structure and its applications. *Chemical Reviews* **1991**, *91*, 893–928.

(43) Smyth, C. P.; McAlpine, K. B. The Dipole Moments of Phosgene, Hydrogen Cyanide and Certain Substituted Methanes. *Journal of the American Chemical Society* **1934**, *56*, 1697–1700.

(44) Johnson, S. The NLopt nonlinear-optimization package.

(45) O. Madelung U. Rossler, M. S. Lead sulfide (PbS) crystal structure, lattice parameters, thermal expansion. *SpringerMaterials*.

(46) Haynie, D. Biological Thermodynamics. **2001**.

(47) Eaves, J.; Loparo, J.; Fecko, C. Hydrogen bonds in liquid water are broken only fleetingly. *PNAS* **2005**.

(48) Box, G.; Cox, D. An analysis of transformations. *Journal of the Royal Statistical Society.* **1964**, *26*, 211–252.